1. In research, it was investigated how the tensile strength of a paper depends on the percentage of the hardwood portion in raw material mixture X_1=Hardwood and (mechanical) scrubbing pressure X_2=pressure during the manufacturing process of paper. Below is a part of the material available in the study. The entire material can be found in dataset paper.txt.

```
     strength hardwood pressure
1       196.6        2      400
2       197.7        2      500
3       199.8        2      650
4       198.4        2      400
.
35      197.8        8      500
36      199.8        8      650
```

Denote explanatory variables as X_1=hardwood and X_2=pressure. Consider modelling the response variable Y=strength by following two different models:

$$\begin{aligned} \mathcal{M}_1: &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i, \\ \mathcal{M}_{1|2}: &\quad Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i, \end{aligned}$$

where in each model the random error term ε_i is assumed to follow normal distribution $\varepsilon_i \sim N(0, \sigma^2)$.

(a) Under the model $\mathcal{M}_{1|2}$, calculate the maximum likelihood estimate for the parameter β_2.

(1 point)

(b) Under the model $\mathcal{M}_{1|2}$, find the restricted maximum likelihood estimate, i.e., an unbiased estimate $\tilde{\sigma}^2$ for the the variance parameter σ^2. (1 point)

(c) Under the model $\mathcal{M}_{1|2}$, calculate the fitted value $\hat{\mu}_1$ for the first observation $i = 1$ in the data set.

(1 point)

(d) Under the model $\mathcal{M}_{1|2}$, calculate maximum likelihood estimate for the expected value μ_{i_*}, when $\mathbf{x}_{i_*1} = 7$ and $x_{i_*2} = 550$. (1 point)

(e) Under the model $\mathcal{M}_{1|2}$, calculate the 80% prediction interval for the new observation Y_{i_*}, when $x_{i_*1} = 7$ and $x_{i_*2} = 550$.

(1 point)

(f) Consider the following hypotheses

$$\begin{aligned} H_0 &: \text{Model } \mathcal{M}_1 \text{ is the true model}, \\ H_1 &: \text{Model } \mathcal{M}_{1|2} \text{ is the true model}. \end{aligned}$$

Select the appropriate test statistic to test the above hypotheses. Calculate the value of the test statistic. (1 point)

2. The dream of every Karate Kid is to find such a punching board (makiwara board) that will withstand the blows but which would not be so rigid or hard that training would then harm hands. The makiwara board can be made in different kinds of wood. In study, it was examined how much a makiwara board bends (in millimeters) of the force of the strike in different tree species. The makiwara boards used in study were made in two different ways. Dataset is given in file makiwaraboard.txt.

```
       WoodType BoardType Deflection
1             1         1      144.3
2             1         1      125.9
3             1         1      263.2
4             1         1      114.6
5             1         1      242.5
6             1         1      141.9
.
.
335           4         2       73.3
336           4         2       44.9

Description: Results of experiments measuring deflection (mm) of
makiwara boards of two types (stacked and tapered) and of four
wood types (Cherry, Ash, Fir, and Oak).

Wood Type: 1=Cherry, 2=Ash, 3=Fir, 4=Oak
Board Type: 1=Stacked, 2=Tapered

Source: P.K. Smith, T. Niiler, and P.W. McCullough (2010). "Evaluating
Makiwara Punching Board Performance," Journal of Asian Martial Arts,
Vol 19, #2, pp. 34-45.
```

Denote explanatory variables as X_1=WoodType and X_2=BoardType. Consider modelling the response variable Y=Deflection by following two different models:

$$
\begin{aligned}
\mathcal{M}_{1|2}: \quad & Y_i \sim N(\mu_{jh}, \sigma^2), \\
& \mu_{jh} = \beta_0 + \beta_j + \alpha_h, \\
\mathcal{M}_{12}: \quad & Y_i \sim N(\mu_{jh}, \sigma^2) \\
& \mu_{jh} = \beta_0 + \beta_j + \alpha_h + \gamma_{jh},
\end{aligned}
$$

where index j is related to the categories of the variable X_1=WoodType and index h is related to the categories of the variable X_2=BoardType.

(a) Under the model $\mathcal{M}_{1|2}$, calculate the maximum likelihood estimate for the expected value μ_{jh}, when the explanatory variables X_1, X_2 are set on values

$$X_1 = \mathsf{Oak} = 4,$$
$$X_2 = \mathsf{Tapered} = 2.$$

That is, find the maximum likelihood estimate for the expected value μ_{42}.

(1 point)

(b) Let us assume that the model $\mathcal{M}_{1|2}$ fits sufficiently enough to the given data set. In which tree species the estimate of the expected value μ_{jh} is in highest level?

i. Cherry,
ii. Ash,
iii. Fir,
iv. Oak.

(1 point)

(c) Consider the following hypotheses

$$H_0 : \text{Model } \mathcal{M}_{1|2} \text{ is the true model},$$
$$H_1 : \text{Model } \mathcal{M}_{12} \text{ is the true model}.$$

Use the Wald statistic

$$W = \frac{(\mathbf{K}'\hat{\boldsymbol{\beta}})'(\tilde{\sigma}^2\mathbf{K}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{K})^{-1}\mathbf{K}'\hat{\boldsymbol{\beta}}}{q}$$
$$= \frac{(\mathbf{K}'\hat{\boldsymbol{\beta}})'\left(\widetilde{\mathrm{Cov}}(\mathbf{K}'\hat{\boldsymbol{\beta}})\right)^{-1}\mathbf{K}'\hat{\boldsymbol{\beta}}}{q} \sim F_{q,n-(p+1)},$$

to test the above hypotheses. Construct appropriate matrix $\mathbf{K}$ and then calculate the value of the test statistic.

(2 points)

(d) Under the model $\mathcal{M}_{12}$, calculate the maximum likelihood estimate for the difference $\mu_{42} - \mu_{11}$. (1 point)

(e) Under the model $\mathcal{M}_{12}$, consider the predictive effect size $Y_{2f} - Y_{1f}$ in situation where the explanatory variables are changed from the values

$$X_1 = \mathsf{Cherry} = 1,$$
$$X_2 = \mathsf{Stacked} = 1.$$

to the values

$$X_1 = \mathsf{Oak} = 4,$$
$$X_2 = \mathsf{Tapered} = 2.$$

Either construct appropriate 80 % prediction interval or test the hypotheses

$$H_0 : y_{1f} = y_{2f},$$
$$H_1 : y_{1f} \neq y_{2f}.$$

How do you interpret your prediction interval or hypothesis testing? Note that this question is heavily related to question (d).

(1 point)

3. (a) Consider the following small data, where X_1 and X_2 are categorical explanatory variables both having class values $\{a, b, c\}$.

	X1	X2	Y
1	a	a	89.0
2	a	b	104.4
3	a	c	97.9
4	b	a	105.5
5	b	b	103.7
6	b	c	104.6
7	c	a	103.3
8	c	b	93.5
9	c	c	102.3

Consider modelling the response variable Y by the following linear model:

$$\mathcal{M}_{1|2}: \quad Y_i = \beta_0 + \beta_j + \alpha_h + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2).$$

The model $\mathcal{M}_{1|2}$ can be written in matrix form as

$$\mathcal{M}_{1|2}: \ \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \mathrm{Cov}(\mathbf{y}) = \sigma^2\mathbf{I}.$$

Write in details what kind forms the model matrix $\mathbf{X}$ and parameter vector $\boldsymbol{\beta}$ have in case of given data is modelled by the model $\mathcal{M}_{1|2}$.

(2 points)

(b) Consider the linear model

$$\mathcal{M}: \ \mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}).$$

Let $\hat{\boldsymbol{\beta}}$ be the maximum likelihood estimator for the parameters $\boldsymbol{\beta}$. Calculate the expected value $\mathrm{E}(\mathbf{X}\hat{\boldsymbol{\beta}})$ and the covariance matrix $\mathrm{Cov}(\mathbf{X}\hat{\boldsymbol{\beta}})$.

(2 points)

(c) Consider the linear model with a new observation

$$\begin{pmatrix} \mathbf{y} \\ Y_f \end{pmatrix} \sim N\left[\begin{pmatrix} \mathbf{X}\boldsymbol{\beta} \\ \mathbf{x}_f'\boldsymbol{\beta} \end{pmatrix}, \sigma^2 \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0}' & 1 \end{pmatrix}\right].$$

The maximum likelihood predictor (the best linear unbiased predictor) for the new observation is

$$\hat{Y}_f = \hat{\mu}_f = \mathbf{x}_f'\hat{\boldsymbol{\beta}} = \mathbf{x}_f'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Show that prediction error $e_f = Y_f - \hat{Y}_f$ has the variance

$$\mathrm{Var}(e_f) = \sigma^2(1 + \mathbf{x}_f'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_f).$$

(2 points)